Data Value Analysis for Predicting Insider Threat Risk using a Bayesian Inference Network

Emily Hsieh
Haystax Technology
ehsieh@haystax.com

Julie Boxwell Ard
Haystax Technology
jard@haystax.com

Ginny Boone
Haystax Technology
gboone@haystax.com

Abstract

Data has high value when it makes a large difference in the estimation of insider threat risk. However, there is limited research on approaches to measure the value of data applied to a Bayesian inference network used to predict insider threats. This paper proposes a methodology, the Node Importance Test (NIT), to represent the impact of a given node on the determination of an individual's riskiness. Experiments illustrate how nodes influenced by various data sources have different ranges of impact on insider threat risk prediction. On average, nodes influenced by user activity monitoring (UAM)-file, and -device data impact risk scores the most. Nodes influenced by employment, medical, and legal data impact risk scores the least. In conclusion, the results show that UAM-file data as a whole are more "valuable" than UAM-login/logout data. These conclusions could reasonably be used as grounds to prioritize acquisition of one type of data over another.

1. Introduction

The insider threat problem is an established and active field of research as described by Azaria et al. [1] with recent government mandates that have reinvigorated the field [2, 3, 4]. A variety of analytical approaches have been proposed for insider threat detection, including Bayesian Networks [5, 6] and traditional machine learning [7].

Insider threat detection can leverage both technical and non-technical data types. These data types include behavioral, network-based, criminal, host-based, financial (both business and personal), user activity monitoring (UAM), publicly available, incident report-based, and human resources (HR) data. Within an organization, these disparate data sources are often owned by various departments that must collaborate to support an effective insider threat program.

In our experience, organizations generally begin by using a subset of the data available to the organization at large. This is due to the costs associated with obtaining new data, the infrastructure necessary for data connection and storage, and the personnel who maintain data flow to the detection tool. Starting with a small-scale proof-of-concept effort allows an insider threat manager to demonstrate incremental successes at a lower initial cost, gain momentum, and get buy-in from other organizational departments so that the monitoring program can expand in later phases.

The question naturally arises as to which data to start with. Instead of an opportunistic approach to data use, we developed this methodology to advise organizations on how to prioritize data procurement for their insider threat programs. We present this exposition of our methodology to measure the value of data for motivating data prioritization and acquisition for insider threat detection.

The rest of this paper is organized as follows. Section 2, Related Work, presents a literature review of related research. Section 3 presents the details of the Bayesian network model that was developed using information from subject matter experts (SMEs), research results, customer feedback, and industry best practices. Section 4 presents the assumption and limitations of the proposed methodology. Section 5 presents the details of the proposed methodology, including: an overview of the model and how it was developed; and an introduction to the Node Importance Test (NIT) used to interrogate the model, which is the basis of this data value analysis. Section 6 presents a general discussion of results, including: the categorization of the nodes into data sources; a description of the impact metric; and examination of the range of potential impact of each data source. Finally, a conclusion that summarizes important findings is presented in Section 7, including a ranked list of data sources to prioritize acquisition of data for insider threat prediction.

2. Related Work

Extensive work has been done to define and evaluate different approaches to predicting insider threat risk. Azaria et al. presented a broad, multidisciplinary survey of insider threat that captured methods from various domains including computer science, psychology, criminology, and security [1]. They

examined behavioral analyses that encompassed understanding exfiltration behaviors from a human behavioral perspective as well as supervised machine learning, and they developed several Behavior Analysis of Insider Threat (BAIT) algorithms to show that they can produce realistic recall of risky insider threat behavior [1]. However, comparing the value of the data used in these methods and algorithms was not considered in these analyses.

Schultz [8] and Wood [9] defined various behavioral indicators that predict insider threat. Some common aspects of people who pose a threat are preparatory behavior, meaningful errors, correlated usage patterns, verbal behavior, personality traits, and skills needed to attack a system. Additionally, psychological and social theorists have presented approaches to insider threat detection using Bayesian network computational approaches [5]. While behavioral indicators and computational approaches are critical to insider threat detection, these studies did not measure or differentiate between the impacts of each piece of information. We describe in Section 3 what distinguishes our Bayesian network approach to modeling insider threat from existing approaches.

Glasser and Lindauer examined the use of synthetic data to conduct insider threat research in response to the realities and difficulties of obtaining suitable data for such research [10]. While they developed criteria for realism and examined pragmatic methods for generating insider threat data, a discussion of the value of various data types did not arise.

Miller and Mork studied methods for selecting key data for specific investigations in order to support large integrated datasets using big data [11]. The focus of their work examined a data value chain, which is the framework used to manage data holistically from initial capture to decision making. While the analysis examined integrated datasets, it did not focus on comparing datasets to each other.

Hubbard and Seiersen comprehensively lay out methods for measuring concepts relevant to cybersecurity risk [12]. They discuss the economic value of information relevant for making risk management decisions, particularly concerning the purchase and implementation costs associated with various cybersecurity controls to mitigate the risk of specific types of cybersecurity incidents. The uncertainties they consider involve the probability that different cybersecurity events will occur, and the data they seek will improve the fidelity of those predictions. While some insider threats are cybersecurity incidents, in this context the likelihood of a *specific cybersecurity incident*, such as an exfiltration of a database containing intellectual property, is a relevant data point. This differs from our approach because the data sources we consider are those available internally to the organization to identify *specific individuals* that are likely to pose an insider threat risk.

There have been many efforts to define insider threats and examine various aspects of the problem. Bishop and Gates developed a definition of the insider threat [13], while Probst et al discuss various definitions and explore several organizational challenges for insider threat detection [14]. Bishop et al. established a framework for monitoring priorities based on information about the insiders and the assets they can access [15].

What distinguishes our approach from the examples discussed above is that there are yet to be analyses on measuring the impact and value of discrete data sources used as inputs for these models in the insider threat domain. We do not attempt to measure the economic cost of obtaining or maintaining the data sources discussed in this paper. The results we present show the relative value of using specific data sources for insider threat monitoring. This value is similar to the *benefit* of using a particular data source in the context of a cost-benefit analysis.

When applying these approaches to insider threat detection in practice, organizations must consider data priority. Obtaining data from different sources incurs financial, infrastructure/support, and liability costs. Therefore, understanding the value each data source brings to the program is critical to its overall success within the organization.

3. Using Bayesian Inference Networks for Insider Threat Risk Modeling

The insider threat domain is characterized by heterogeneous data, but except in rare dramatic cases, the available data are often individually weak indicators of risk. Further, unlike other data-intensive problems, for which deep learning is a solution, there is usually not a labeled data set in the insider threat space that can be used as 'ground truth' for learning.

However, there are human experts who can reason from the evidence available and assess risks that individuals are an insider threat. Unfortunately, human experts quickly become overwhelmed by the sheer volume of data.

3.1. Bayesian Networks for Insider Threat

A Bayesian network provides a way to capture the knowledge that is used by a human analyst to reason from incomplete, inconsistent, and/or heterogeneous data to make an assessment of risk. Applying streams of data as evidence to the Bayesian network allows the

automated assessment of large populations to identify insider risks.

A Bayesian network is a probabilistic model with a structure expressed as a directed acyclic graph: nodes represent random variables (in our case discrete binary random variables with two states: true and false); and arcs represent statistical relationships between variables. Each node in the graph also has a local probability distribution. For nodes without parents, the local probability distribution is a prior distribution for the states of a node. For other nodes, the local probability distribution is a conditional distribution for the states of the node given its parents.

We use the terms node, concept, and hypothesis, interchangeably throughout this paper.

3.2. The Haystax Model

We developed a methodology for defining a class of Bayesian networks with binary nodes that does not require extensive knowledge engineering to specify all of the local probability distributions [16]. The resulting Bayesian network contains a target random variable which is the primary hypothesis of interest, '*IsInsiderThreatConcern*', and 127 supporting hypotheses and indicators that are statistically related to the target hypothesis.

To develop the model, we initially defined high level hypotheses that are typically fairly abstract and for which we do not expect to have direct observations. We extended the model by defining supporting hypothesis and indicators that inform the higher-level hypotheses, and we continued to extend the model with more and more detailed indicators until we had random variables that we expect we will (or may) have direct evidence for.

While evidence can be applied to any node in a Bayesian network, in practice we identified those nodes for which we expect to have evidence. These evidence nodes were the nodes used in our data value methodology.

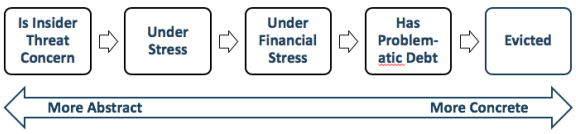

Figure 1. Levels of abstraction in the Haystax model

Our model was constructed using the methodology by Wright et. al., [16] with expertise defined in various documents [17, 18, 19], elicited from SMEs [20, 21, 22], research [23, 24, 25, 26, 27], best practices [10, 28, 29], and case studies [30]. It was refined through assessment of test cases using real and synthetic data.

Suppose that an organization's employees had authorized the company to do a public records check on them. In the diagram depicting a subset of the model shown in Figure 1, you can see that if an organization were to receive publicly available data that someone had been *Evicted*, that this is an indicator that influences *HasProblematicDebt*. If a person has problematic debt, it implies that they are *UnderFinancialStress* which in turn implies that they are *UnderStress*. Furthermore, someone who is under stress is more likely to be an insider threat risk.

The top-level hypothesis node captures the model's belief as to whether an individual represents an insider threat risk to an organization (*IsInsiderThreatConcern)*. From there, 127 additional concepts form the Bayesian inference network that influence this hypothesis.

Nodes can express a positive or negative sentiment and are linked with both positive and negative polarities as well as the strength of the connection [16]. For example, the node *DemonstratesComplianceWithRules* has a positive sentiment. It is strongly linked with negative polarity to the node, *BreaksRules*, which has a negative sentiment.

Target beliefs are assigned to a small number of nodes, and the remainder of the prior beliefs for nodes without target beliefs are calculated using Jeffrey's rule [31] as well as Bayesian inference [32].

Each node is assigned a prior belief that is calculated from selected target beliefs as well as the strength and polarity of the links that connect the network. Suppose we wanted a node to represent the likelihood that an employee is a binge drinker. We would create the node, *BingeDrinker,* and could link it to *AbusesAlcohol*. By applying a target belief, we would set the prior for this *BingeDrinker* node to 0.269 to reflect the probability that an American adult has used alcohol to excess within the last month [33]. In the absence of data that could confirm or deny this assertion, the model has a default belief of 0.269 that a person binge drinks. This initial value would be the prior belief.

Suppose an organization received data during a pre-hire criminal check that an employee was convicted of a driving under the influence (DUI) violation. The Haystax platform would apply this data to the *BingeDrinker* node using the process explained in Section 3.3, and this would result in an increased belief of, say, 0.75. This second value is the posterior belief. The posterior belief of the top-level hypothesis node, *IsInsiderThreatConcern*, estimates the likelihood that an employee is an insider threat risk to the organization. Having constructed the Bayesian network, we have an initial belief for each random variable in the model, including the top-level hypothesis '*IsInsiderThreatConcern*'. This initial belief, before

any evidence is applied, is the *prior belief*. When we apply evidence to the model, the beliefs are updated, consistent with the data, to yield an updated belief, which is the *posterior belief* for each node.

3.3. Applying Data to the Model

Now that we have described the Haystax model, we can discuss how data is applied to it. Several different types of data can indicate a concept—i.e., can be evidence for a concept (random variable). For example, shoplifting is an indicator of *CommitsNonViolentCrime*, but so is embezzlement—with a different severity than shoplifting. To build a model with each node representing every possible data type would result in an unacceptably slow runtime for an operational product. Therefore, this framework allows data types to be applied to nodes using specified ingestion parameters. The parameters include growth and decay half-life, strength, and polarity [34].

During the data ingestion process, data are applied to our insider threat (InT) model at the lowest, most concrete level. Figure 1 illustrates how abstract concepts are broken down into concrete concepts that indicate behaviors relevant to the insider threat domain. These concrete concepts are called "evidence nodes," and we can apply observable data to them.

Growth and decay half-life allow us to increase or decrease the importance of events over time. Specifying a decay half-life enables us to differentiate between a DUI that a person received yesterday from a DUI received 20 years ago. In the same way, we can increase the importance of events over time using a growth half-life. Half-life is specified in days and can be infinite [34]. This half-life is measured against the date of the evaluation, which defaults to the current date, but can be otherwise specified.

We can also distinguish the "quality" of different types and sources of data. Suppose a coworker reports another coworker to HR because the employee recently heard a fellow employee bragging in the break room about shoplifting. This coworker report has a lower quality of "evidence" than, say, a police report for a shoplifting incident. In this way, we can differentiate the two events using different ingestion strengths. Our framework employs 12 levels for ingestion strength, including "absolute" strength [34]. We also can apply an observable event against a model node with reverse polarity if needed. For example, if an individual was promoted, that could be applied as evidence that decreases the belief in the *Demoted* node. An arbitrarily large number of events can be applied to one node.

4. Assumptions and Limitations

4.1. Assumptions

We begin with the assumption that the Haystax InT model, its framework, concept network, connection strengths, and observable event ingestion parameters, are a valid representation of the complex factors that influence reasoning about insider threats. We assert that it represents an accurate assessment of insider threat risk because it was built using the widely accepted policies, case studies, research, SME interviews, and best practices referenced above in Section 3.2. It was further validated with customer data and refined based on feedback from customers and external SMEs. The following methodology is general, but the results are specific to this particular version of the insider threat model.

This methodology also assumes that the impact a piece of evidence applied to this model has on an individual's probabilistic risk score is positively correlated with the value of the data that caused the impact. Data that have a larger impact on the risk estimate are therefore considered "more valuable" to this model. In other words, data that can be obtained from a given data source are considered more valuable if the model nodes they are applied to have a higher impact on risk determination. Section 3.3 describes how data are applied to the model.

4.2. Limitations

This methodology applies only to Bayesian inference networks. The analysis we use to assess data source impact and value is specific to a risk assessment process built using Bayesian networks according to the process discussed in Section 3.

Additionally, the Node Importance Test (NIT), further discussed in Section 5.1, returns the ceiling of the importance value for each node it evaluates. As outlined below in Section 5.1, the NIT assumes the likelihood of a concept is absolute, and therefore the output and corresponding impact value represent the maximum impact the concept can have on the overall risk score. As an illustration of this nuance, using the example presented above, evidence of shoplifting would be applied to the node *CommitsNonViolentCrime*, with a weak strength and a short half-life, while a more serious nonviolent crime such as embezzlement would be applied to that same node with a strong ingestion strength and a longer half-life. By design, the impact of either the shoplifting or embezzlement events would be less than absolute on the node *CommitsNonviolentCrime*. Using this methodology, the

NIT assesses the maximum impact that *CommitsNonViolentCrime* can have on overall risk likelihood.

Finally, model nodes are evaluated independently. The impact of a single node by itself is higher than its contribution to the total impact of multiple nodes in conjunction. In practice, an individual's risk score is derived from multiple data events applied to multiple nodes. Therefore, the incremental value of each node when multiple types of data are simultaneously applied to multiple nodes is lower than when nodes are applied independently. Due to the large number of combinations of 99 nodes, analyzing incremental value in a comprehensive manner is difficult to present concisely. When needed, we can measure the incremental value of adding a data source to an existing combination of data sources on a case-by-case basis.

5. Methodology

As described above in Section 3, the Haystax InT model is a whole-person, Bayesian inference network that estimates insider threat risk. The purpose of the model is to determine the likelihood that an individual employed by an organization represents an insider threat concern.

To formulate the methodology that quantified the impact on the model of data from different sources, we designed a test called the Node Importance Test. The NIT measures changes in predictive probabilities associated with each node in the Bayesian network. Results of the NIT, further explained in the following section, represent the impact a given node has on the overall determination of an individual's riskiness.

The NIT methodology is informed by how data is applied to the model. During the data ingestion process, detailed in Section 3, data are ingested into the model at the most concrete level. Therefore, only evidence nodes were included in the analysis.

5.1. The Node Importance Test (NIT)

We developed the NIT to assess the relative importance or impact of the nodes in a Bayesian network. The test generates a quantitative measure of the influence each node has on the highest-level node in the model, *IsInsiderThreatConcern.*

The NIT process is as follows:

1. Apply a single, unique event to each node with an "absolute" strength and an event date that is equal to the evaluation date. This has the effect of setting each node belief equal to 1, its maximum value.
2. Calculate the difference between the posterior belief and the prior belief of the value of the *IsInsiderThreatConcern* node resulting from the application of each "absolute" event to its corresponding node. This value represents the difference in the risk score that each node is responsible for.
3. Convert the difference obtained in step 2 to the "impact" metric, described in the next section. This provides a quantitative measure of the influence that each node has on the risk score.

Note that this test quantifies the influence of the concepts represented by the nodes in the Bayesian network, but it does not necessarily imply that an event applied to a stronger node has a greater impact than an event applied to a weaker node. Events are discussed in more detail in Section 3.3.

5.2. Measuring Impact

When the Haystax platform ingests information and applies it to a Bayesian network, the primary output is a number in the range from 0 to 1, which expresses its belief that an individual represents an insider threat concern to their organization. This is also know as the "risk score." A risk score of 0 indicates that the model believes that particular individual has a null probabilistic chance of being an insider threat concern to their organization. A risk score of 1 indicates that the model believes the individual has an absolute likelihood of representing an insider threat concern.

The results of the NIT are the differences of the prior and posterior beliefs of the highest-level node, *IsInsiderThreatConcern*. Prior and posterior beliefs are explained above in section 3.2. We developed and applied a metric—impact—to better represent the significance of that change in probability on an evidence node to the top-level *IsInsiderThreatConcern* risk score.

We developed the impact transformation metric to compare changes in risk belief rather than a simple belief difference because odds ratios largely determine how beliefs propagate through a Bayesian network. Using the absolute value of the log transformation allows data that influence the *IsInsiderThreatConcern* node with differing polarities to be considered together.

Impact is defined as the absolute value of the difference of the natural logarithms of the ratios of the prior and posterior odds:

$$|\ln(Odds.Posterior) - \ln(Odds.Prior)|$$

Or, more directly:

$$\left|\ln\left(\frac{Belief.Posterior}{1-Belief.Posterior}\right) - \ln\left(\frac{Belief.Prior}{1-Belief.Prior}\right)\right|$$

We will consider an example in order to illustrate the impact metric. If a belief changed from 0.99 to 0.90, we would have:

$$\ln(9) - \ln(99) = -\ln(11)$$

While the belief difference is only -0.09, the odds ratios would change from 99:1 to 9:1. This is a factor of 11, with an impact value of -2.4.

If a belief changed that same amount (-0.09) from 0.54 to 0.45, the odds ratios would change by a factor of only about 0.7, with an impact value of 0.36. Despite the same "change in score," the factor of change, or impact, is greater when the belief is closer to 0 or 1. Data that cause changes closer to the edges of the belief distribution have a higher impact than data that cause changes closer to the center of the distribution.

5.3. Experimental Design

This section presents the experiments we performed to analyze the impact of data sources on the Haystax insider threat risk model. We begin with a description of the different categories of data sources, discuss how nodes were classified, and then explain the analysis process.

To measure the "value" of data sources applied to the insider threat model, each node in the model is classified by potential data source(s) using the criterion: "could evidence from data source X be applied to the current evidence node?" where X is one of 12 data source types listed in Table 1.

A single node could be classified under multiple data sources. For example, the node, *HasBankruptcy*, was classified under both financial and legal data sources because evidence of bankruptcy could potentially be obtained from either source.

This experiment compared the impact of all model nodes at the data source category level. For example, to quantify the value of criminal data compared to financial data, we compared the impact of nodes that could be influenced by criminal data to the impact of nodes that could be influenced by financial data. In this sense, we compared the maximum impact of nodes that could be influenced by evidence from 12 different data sources, described in Table 1.

Table 1. Descriptions of data sources

Data Source	Description
Criminal	Information from an individual's criminal record, such as felonies, misdemeanors, white collar crime, traffic infractions, and domestic violence.
Employment-HR	Information from an individual's employer and human resources records, such as job performance reviews, promotions, written warnings, formal verbal warnings, policy violations, formal complaints, and termination.
Financial	Information about an individual's financial behavior and financial records, such as expense reports, use of a company credit card, credit score, bankruptcy, financial transactions, and trading activity.
Legal	Information from an individual's legal records, such as divorce, custody disputes, restraining orders, domestic violence, tax lien, foreclosure, eviction, and bankruptcy.
Medical	Information from an individual's medical records, such as mental health, physical health, and alcohol and drug abuse.
Badge	Information about an individual's physical access within employer spaces.
UAM-Device	Information about an individual's network device activity, such as attempts to exfiltrate files via a removable device.
UAM-Email	Information about an individual's organizationally-owned email activity and any information that could be revealed through this email communication.
UAM-File	Information about an individual's file activity on organization-owned systems, including attempts to access, use, rename, copy, and delete files.
UAM-LogIn/Out	Information about an individual's login and logout activity on organization-owned systems and resources.
UAM-Print	Information about an individual's activity on organization-owned printers.
UAM-Website	Information about an individual's website activity accessed from organization-owned systems, including website URLs.

The inputs for this NIT were 99 evidence nodes out of the total of 128 nodes in the Haystax InT model (including the top-level hypothesis node). As explained in Section 3, an evidence node is any node in the Bayesian network with the potential for data to be applied as evidence directly to the node. As described in Section 5.1, the NIT output quantified how an absolute likelihood for a specific node affected the overall *IsInsiderThreatConcern* belief score.

Each node was independently tested and the subsequent change in risk score is solely due to the application of the singular belief that the node is set to "true." In this way, we measured the impact of a single node in isolation. Figure 2 summarizes the experimental process.

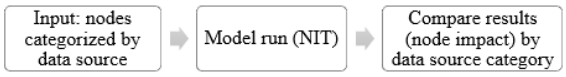

Figure 2. Data source comparison process

6. Results

This experiment illustrated how nodes influenced by various data sources have different ranges of impact on insider threat risk prediction. Figure 3 depicts the impact of nodes that are influenced by evidence from different data sources. For each data source, the box plot displays the distribution of the maximum impact values for all nodes that can be influenced by that data source.

On average, nodes influenced by UAM-file and -device data impacted the risk score the most. Nodes influenced by employment, medical, and legal data impacted the risk score the least.

Criminal, legal, and employment data sources have the widest range of potential impact. This is likely due to the broad range of seriousness of criminal, legal, and employment activity. Information from these sources applied at absolute strength could have a significant impact on an individual's risk score while other data from the same source could have a minimal impact.

In contrast, UAM-login/logout data has a small range of potential node impact. Additionally, the maximum impact of UAM-login/logout data is less than the minimum impact of UAM-file data. Comparing these examples, we conclude that UAM-file data as a whole is more "valuable" than UAM-login/logout data. These conclusions could reasonably be used as grounds to prioritize acquisition of one type of data over the other.

7. Conclusions

The Haystax insider threat model is a whole-person risk model. Because it uses multiple sources of data to assess an individual's risk, this Data Value Analysis sought to understand how data from each source impacted the model.

The experiment illustrated the differential impact that changes in node likelihood have on the InT model.

Figure 3. Box plots of maximum node impact by potential data source

On average, nodes influenced by UAM-file and -device data impacted the model the most, while nodes influenced by employment, medical, and legal information impacted the model the least.

We developed this methodology to assess the impact of data sources on a Bayesian inference network risk model in order to measure the relative value of different data sources and motivate prioritizing acquisition of more valuable data sets.

8. Future Work

To address the experimental limitations of the NIT, which restricted results to a node's "maximum potential impact", future experiments will test variations in the ingestion parameters of the Haystax insider threat model. By applying observable events with different ingestion parameters, such as growth or decay half-life and ingestion strength in a notional "Event Importance Test," we can produce a more granular and nuanced understanding of how data may impact the model.

Furthermore, this methodology can be compared to traditional statistical measures of value and sensitivity. For example, we can compare the ranked results from our NIT to measures of Entropy and Mutual Information commonly used in sensitivity analyses.

We also seek to validate our approach empirically through feedback from customers. To augment our understanding of the impact that different data types have on overall insider threat risk in isolation, we can test both combinations and incremental impact of additions. Drawing on our experience, we can test likely combinations of data sources to identify the optimal data a customer could use to start an insider threat program. From there, we can measure the value that additional data sources would bring when a customer is ready to expand an insider threat program.

9. References

[1] A. Azaria, A. Richardson, S. Kraus, and V. S. Subrahmanian, "Behavioral Analysis of Insider Threat: A Survey and Bootstrapped Prediction in Imbalanced Data", IEEE Transactions on Computational Social Systems, IEEE, vol. 2, no. 2, pp. 135-155, 2014.

[2] S. Miller and J. Trotman, "Insider Threats in Finance and Insurance (part 4 of 9: Insider Threats Across Industry Sectors)", Insider Threat Blog, Software Engineering Institute at Carnegie Mellon University, December 2018.

[3] U.S. Department of Defense, "National Industrial Security Program Operating Manual", DoD, May 2016.

[4] The White House, "Executive Order 13587 – Structural Reforms to Improve the Security of Classified Networks and the Responsible Sharing and Safeguarding of Classified Information", The White House of President Barack Obama, Office of the Press Secretary, October 2011.

[5] T. Read, S. Maclsaac, V. Corsi, R. Campbell, T. Crawford, C. Davenport, B. Denson, D. McGarvey, D. Thomas, "An Assessment of Data Analytics Techniques for Insider Threat Programs", Intelligence Insights, INSA, July 2018.

[6] Software Engineering Institute, "Analytical Approaches to Detect Insider Threats", Software Engineering Institute, Carnegie Mellon University, Pittsburgh, PA, December 2015.

[7] Emaasit, D. and Johnson, M., 2018. Capturing Structure Implicitly from Time-Series having Limited Data. *arXiv preprint arXiv:1803.05867*.

[8] E. E. Schultz, "A Framework for Understanding and Predicting Insider Attacks," Computers & Security, vol. 21, no. 6, pp. 526–531, 2002.

[9] B. Wood, "An Insider Threat Model for Adversary Simulation," SRI International, Research on Mitigating the Insider Threat to Information Systems, vol. 2, pp. 1–3, 2000.

[10] J. Glasser and B. Lindauer, "Bridging the Gap: A Pragmatic Approach to Generating Insider Threat Data", 2013 IEEE Secutiy and Privacy Workshops, IEEE, San Francisco, CA, USA, 2013.

[11] H. G. Miller and P. Mork, "From Data to Decisions: A Value Chain for Big Data", IT Professional, vol. 15, no. 01, pp. 57-59, 2013.

[12] D. Hubbard and R. Seiersen, How to Measure Anything in Cybersecurity Risk, 1st ed., pp. 189-195, Wiley, 2016.

[13] M. Bishop and C. Gates, "Defining the Insider Threat," *Proceedings of the Cyber Security and Information Intelligence Research Workshop* article 15 (May 2008).

[14] Probst C.W., Hunker J., Gollmann D., Bishop M. (2010) Aspects of Insider Threats. In: Probst C., Hunker J., Gollmann D., Bishop M. (eds) Insider Threats in Cyber Security. Advances in Information Security, vol 49. Springer, Boston, MA.

[15] M. Bishop, C. Gates, D. Frincke, and F. Greitzer, "AZALIA: an A to Z Assessment of the Likelihood of Insider Attack," *Proceedings of the 2009 IEEE International Conference on Technologies for Homeland Security* pp. 385–392 (May 2009).

[16] E. Wright, R. Schrag, R. Kerr, and B. Ware, "Automating the Construction of Indicator-Hypothesis Bayesian Networks from Qualitative Specifications", Haystax Technology technical reports, 2015.

[17] Defense Personnel and Security Research Center and Defense Manpower Data Center, “Adjudicative Desk Reference Version 4”, U.S. Department of Defense, March 2014.

[18] Federal Deposit Insurance Corporation, “Risk Management Manual of Examination Policies”, Federal Deposit Insurance Corporation (FDIC), June 2019.

[19] Office of the Director of National Intelligence, “Security Executive Agent Directive-4 (SEAD-4) National Adjudicative Guidelines”, U.S. Department of State, June 2017.

[20] Coggins, Margaret. Subject Matter Expert Personal Interview. Haystax meeting with fraud SMEs, September 2017.

[21] Hrdlicka, Angela. Subject Matter Expert personal interview. Haystax meeting with fraud SMEs, September 2017.

[22] Page, DC. Subject Matter Expert personal interview. Haystax meeting with fraud SMEs, September 2017.

[23] E.T. Axelrad, P. J. Sticha, O. Brdiczka and J. Shen, “A Bayesian Network Model for Predicting Insider Threats”, 2013 IEEE Security and Privacy Workshops, 2013.

[24] F. L. Greitzer and D. A. Frincke, “Combining Traditional Cyber Security Audit Data with Psychosocial Data: Towards Predictive Modeling for Insider Threat Mitigation,” in Insider Threats in Cyber Security. Springer, 2010, pp. 85–113.

[25] F.L. Greitzer, L.J. Kangas, C.F. Noonan, and A.C. Dalton, “Identifying at-Risk Employees: A Behavioral Model for Predicting Potential Insider Threats”, Pacific Northwest National Laboratory, Prepared for U.S. Department of Energy, Richland, WA, September 2010.

[26] J. R. C. Nurse, O. Buckley, P. A. Legg, M. Goldsmith, S. Creese, G. R. T. Wright, and M. Whitty, “Understanding Insider Threat: A Framework for Characterising Attacks”, 2014 IEEE Security and Privacy Workshops, 2014.

[27] E. Shaw and L. Seller, “Application of the Critical-Path Method to Evaluate Insider Risks”, Studies in Intelligence, Internal Security and Counterintelligence at U.S. Central Intelligence Agency, vol. 59, no. 2. June 2015.

[28] A. P. Moore, D. Cappelli, and R. F. Trzeciak, “The ‘Big Picture’ of Insider IT Sabotage Across U.S. Critical Infrastructures”, Software Engineering Institute, Carnegie Mellon University, Pittsburgh, PA, May 2008.

[29] A. P. Moore, D. McIntire, D. Mundie, and D. Zubrow, “Justification of a Pattern for Detecting Intellectual Property Theft by Departing Insiders”, Software Engineering Institute, Carnegie Mellon University, Pittsburgh, PA, March 2013.

[30] PERSEREC, “Espionage and Other Compromises of National Security”, PERSEREC, Defense Personnel Security Research Center, Monterey, CA, November 2009.

[31] Jeffrey, R. The Logic of Decision 2nd Edition, University of Chicago Press, Chicago, 1983.

[32] R. Schrag, E. Wright, R. Kerr, and R. Johnson, “Target Beliefs for SME-Oriented, Bayesian Network-Based Modeling”, 13th Annual Bayesian Modeling Applications Workshop, 2016.

[33] National Institute of Alcohol Abuse and Alcoholism, “Alcohol facts and statistics”, National Institute of Alcohol Abuse and Alcoholism, National Institutes of Health, U.S. Department of Health and Human Services, August 2018.

[34] Robert Schrag, Edward Wright, Robert Kerr, and Bryan Ware, “Processing Events in Probabilis%c Risk Assessment,” *9th Interna1onal Conference on Seman1c Technologies for Intelligence, Defense, and Security (STIDS),* 2014.